Winning Space Race with Data Science

Minglu Zeng
10/14/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory and Preparing Data
  - Exploratory Data Analysis: with sql, with Data Visualization, with Folium, with Plotly Dash
  - Machine Learning Prediction: Logistic Regression, Support Vector Machine, Decision Tree, KNN
- Summary of all results
  - Exploratory Data Analysis result
  - Predictive Analytics result using classification models

Introduction

- Project background and context

According to data released by SpaceX, the Falcon9 rocket costs $62million while other companies often spend more than $165million for a launch. The reason SpaceX can be able to go so much lower is because it can reuse the first stage rocket. If we can calculate the cost of launch by predicting whether the first stage will land, other companies can use that information. In this project, we can use machine learning to predict whether the first stage of SpaceX will land or not

- Problems you want to find answers

➢ Which factors determine the successful rate of landing?

➢ What should be the operating conditions of a successful landing?

Section 1

Methodology

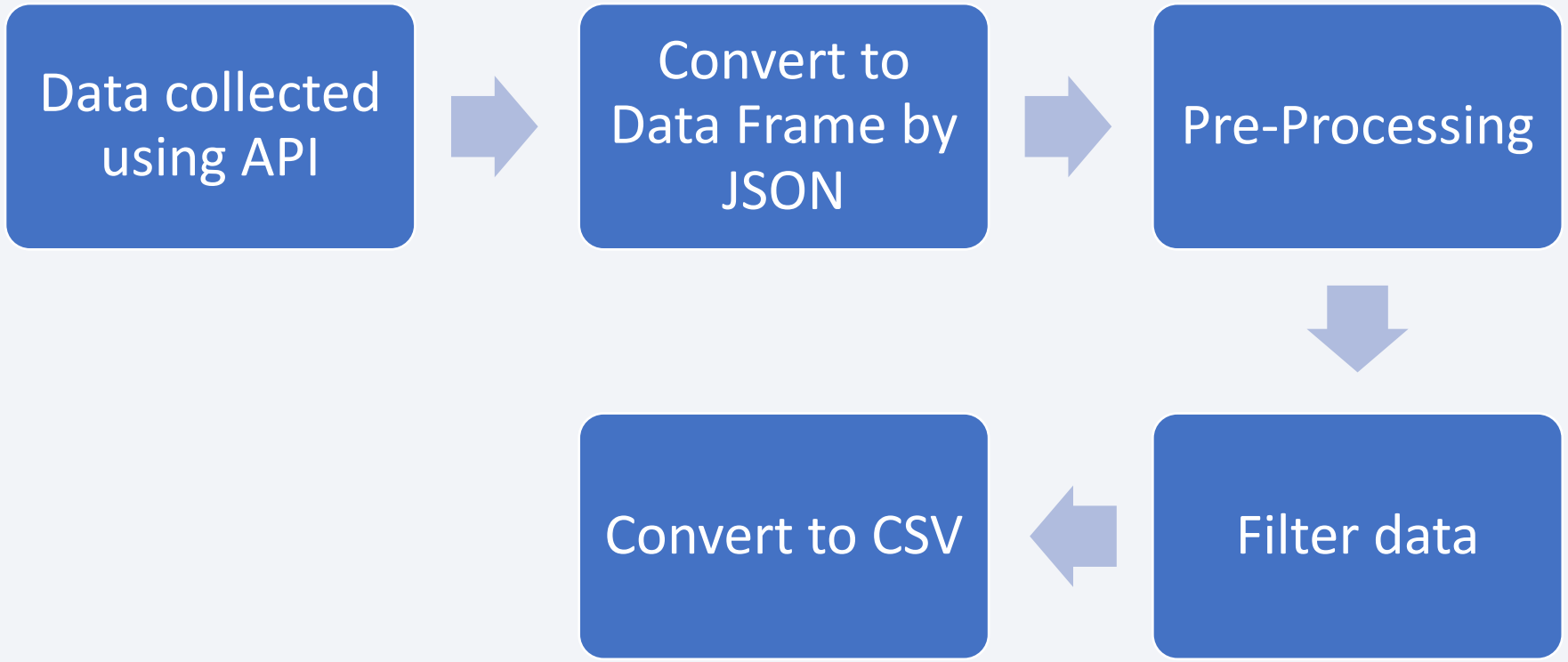
Methodology

Executive Summary

- Data collection methodology:
  - How data was collected with API and Web Scrapping
- Perform data wrangling
  - How data was processed with one hot encoded
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Scatter plot, line plot and bar plot
- Perform interactive visual analytics using Folium and Plotly Dash
  - Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

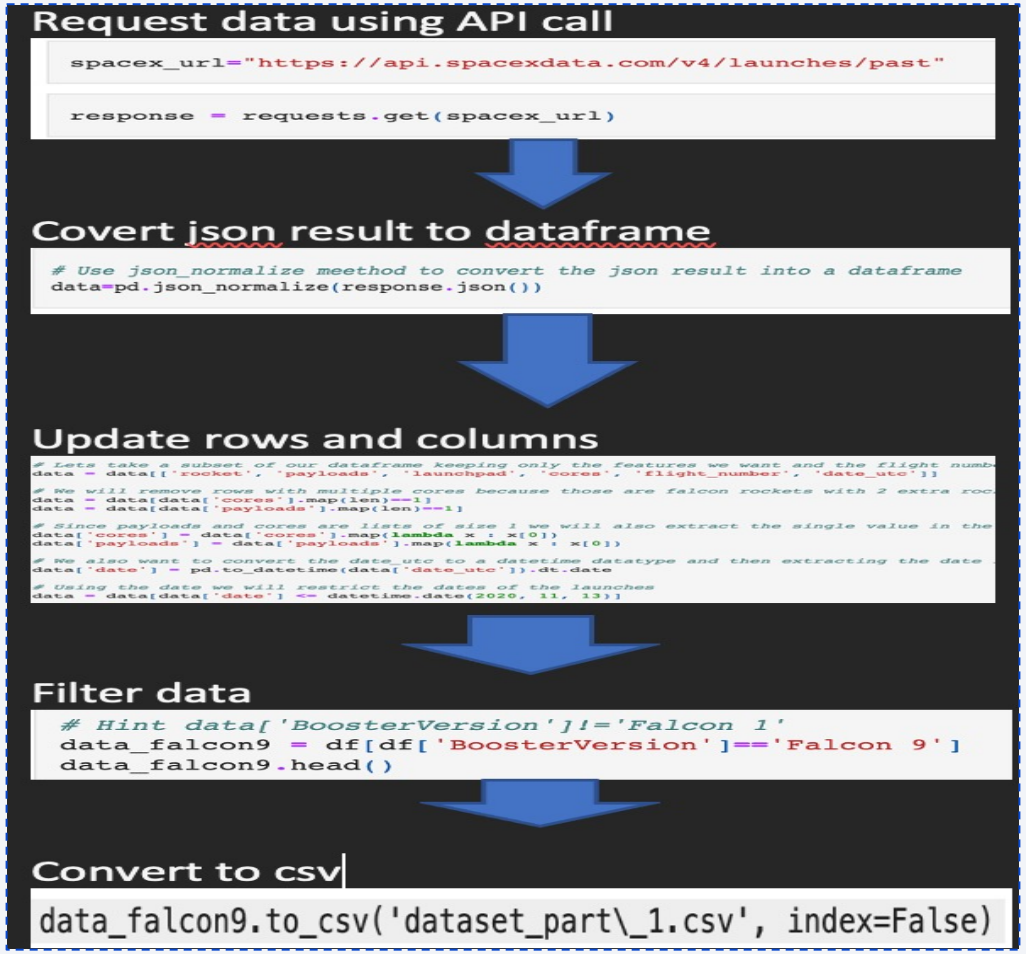
Data sets are collected using the API call from SpaceX Websites.



Data Collection – SpaceX API

Data collection using API call

- https://github.com/mingluzeng/spaceX_assignment/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Falcon 9 historical launch records collected by web scraping with BeautifulSoup
- https://github.com/mingluzeng/spaceX_assignment/blob/main/jupyter-labs-webscraping.ipynb

1.Get responses from HTML

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url).text
```

2.Create Beautiful Soup Object

```
# Use BeautifulSoup() to create a BeautifulSoup
soup = BeautifulSoup(response, 'html.parser')
```

3.Finding tables

```
# Use the find_all function in the Beautifu
# Assign the result to a list called `html_
html_tables = soup.find_all("table")
print(html_tables)
```

4.Getting columns name

```
column_names = []
# Apply find_all() function with `th` element on fir
# Iterate each th element and apply the provided ext
# Append the Non-empty column name (`if name is not
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
        name = extract_column_from_header(temp[x])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

5.Creating dictionary

```
launch_dict= dict.fromkeys(column_names)
# Remove an irrelvant column
del launch_dict['Date and time ( )']
# Let's initial the launch_dict with each va
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6.Appending data to keys

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
    # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corresponding to launch a number
```

7.Coverting dictionaries to data frames

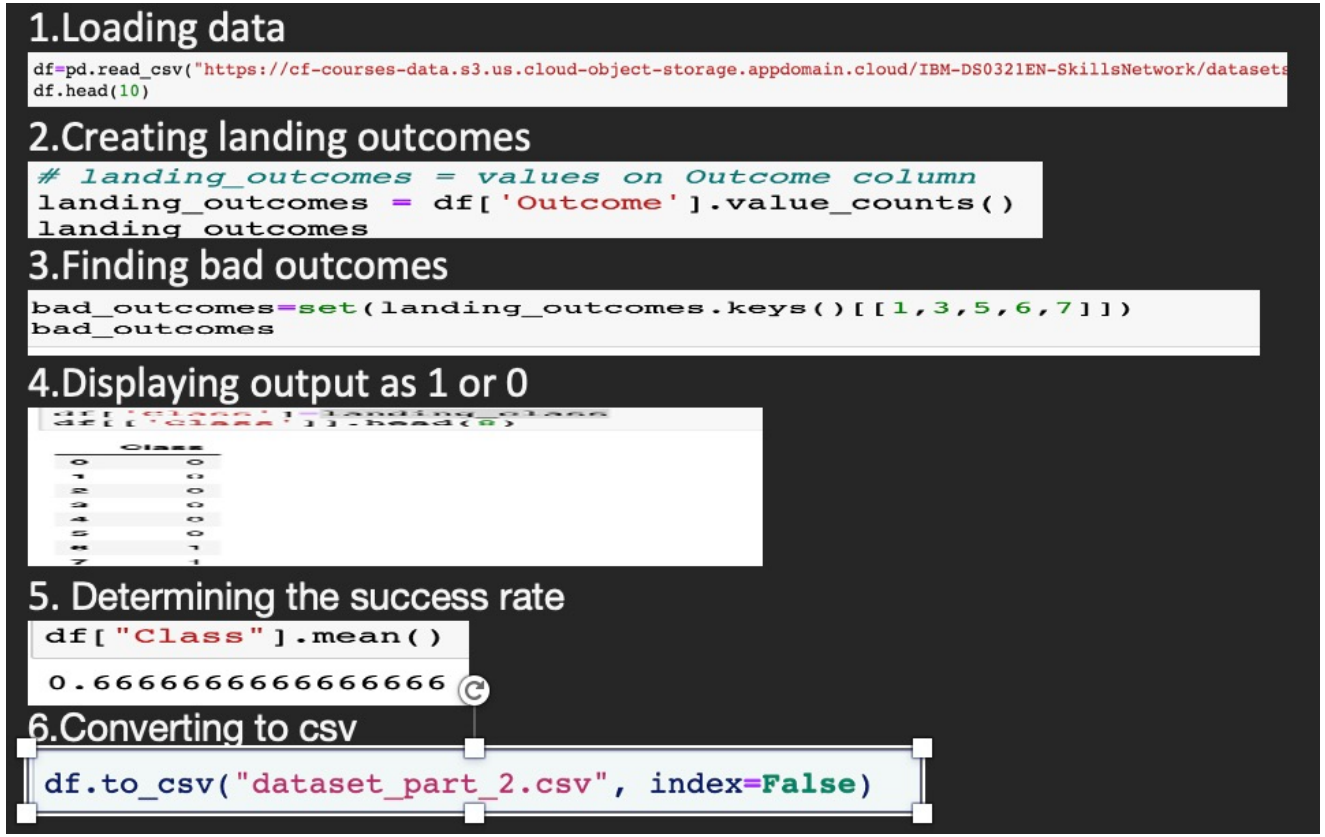
```
df=pd.DataFrame(launch_dict)
```

8.Converting data frames to csv file

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Data Wrangling

- Performed Exploratory data analysis
- Calculated the number of launches of each site, orbit occurrence and mission
- Created class label to calculate successful rate
- https://github.com/mingluzeng/spaceX_assignment/blob/main/labs-jupyter-spacex-Data wrangling.ipynb



EDA with Data Visualization

https://github.com/mingluzeng/spaceX_assignment/blob/main/jupyter-labs-eda-dataviz.ipynb

- **We visualized relationship between successful rate and Launch Site, the relationship between the number of flights and launch sites, the relationship between Payload and Launch Site, the relationship between number of flights and Orbit type by scatter plot chat**
- **We visualized the relationship between success rate of each orbit type by bar chat**
- **We observed launch success yearly trend by line chart**

EDA with SQL

the SQL queries I performed

- **The names of the unique launch sites in the space mission**
- **5 records where launch sites begin with the string 'CCA'**
- **The total payload mass carried by boosters launched by NASA (CRS)**
- **Average payload mass carried by booster version F9 v1.1**
- **The date when the first successful landing outcome in ground pad was acheived.**
- **The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **The total number of successful and failure mission outcomes**
- **The booster_versions which have carried the maximum payload mass**
- **The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015**
- **The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order**

https://github.com/mingluzeng/spaceX_assignment/blob/main/Spacex (1).ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a folium map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

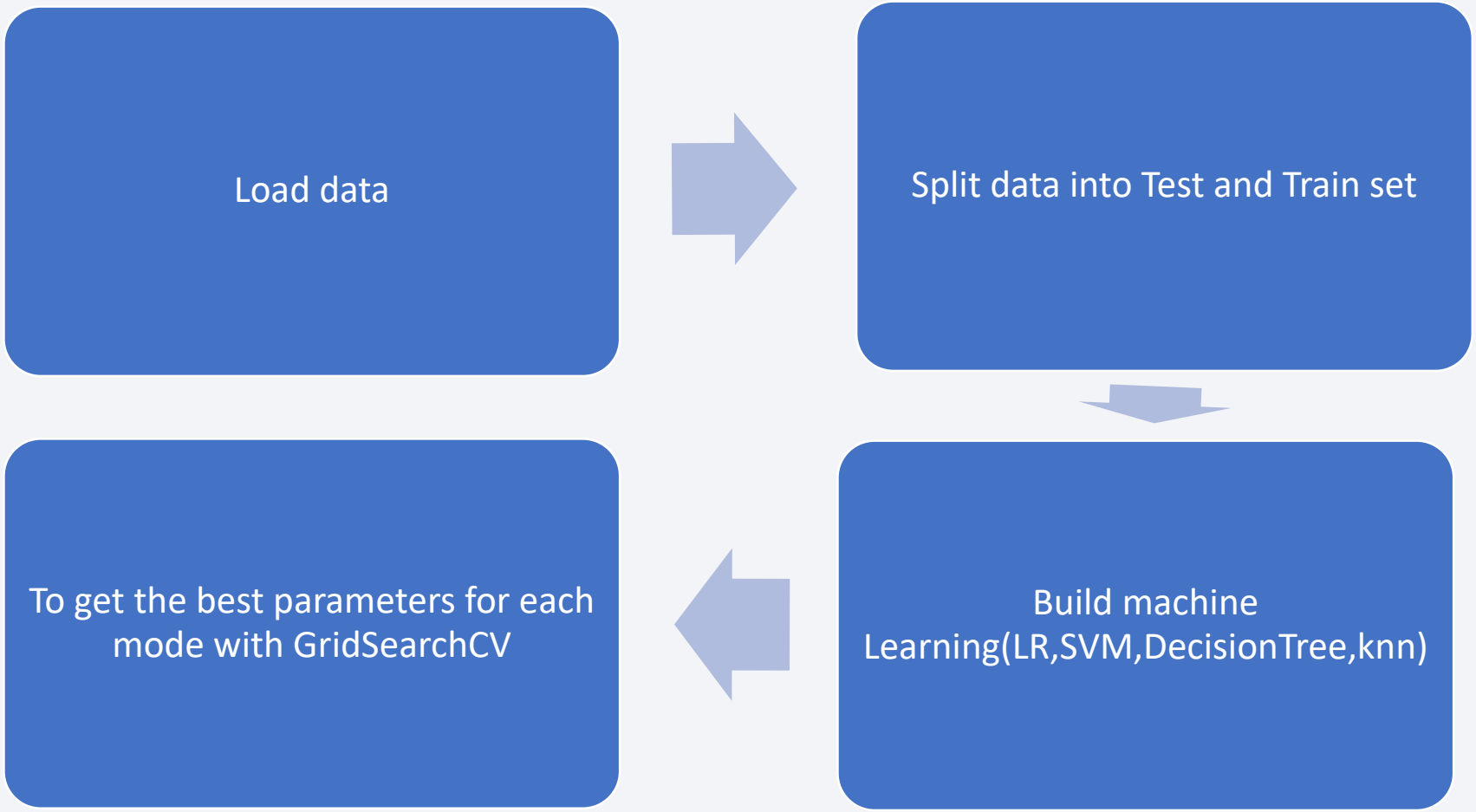
https://github.com/mingluzeng/spaceX_assignment/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We displayed the total launches by a certain sites by pie chart
- We plotted scatter graph displayed the relationship with Outcome and Payload Mass (Kg) for the different booster version by plotted scatter.

https://github.com/mingluzeng/spaceX_assignment/blob/main/Ploty_Dash.py

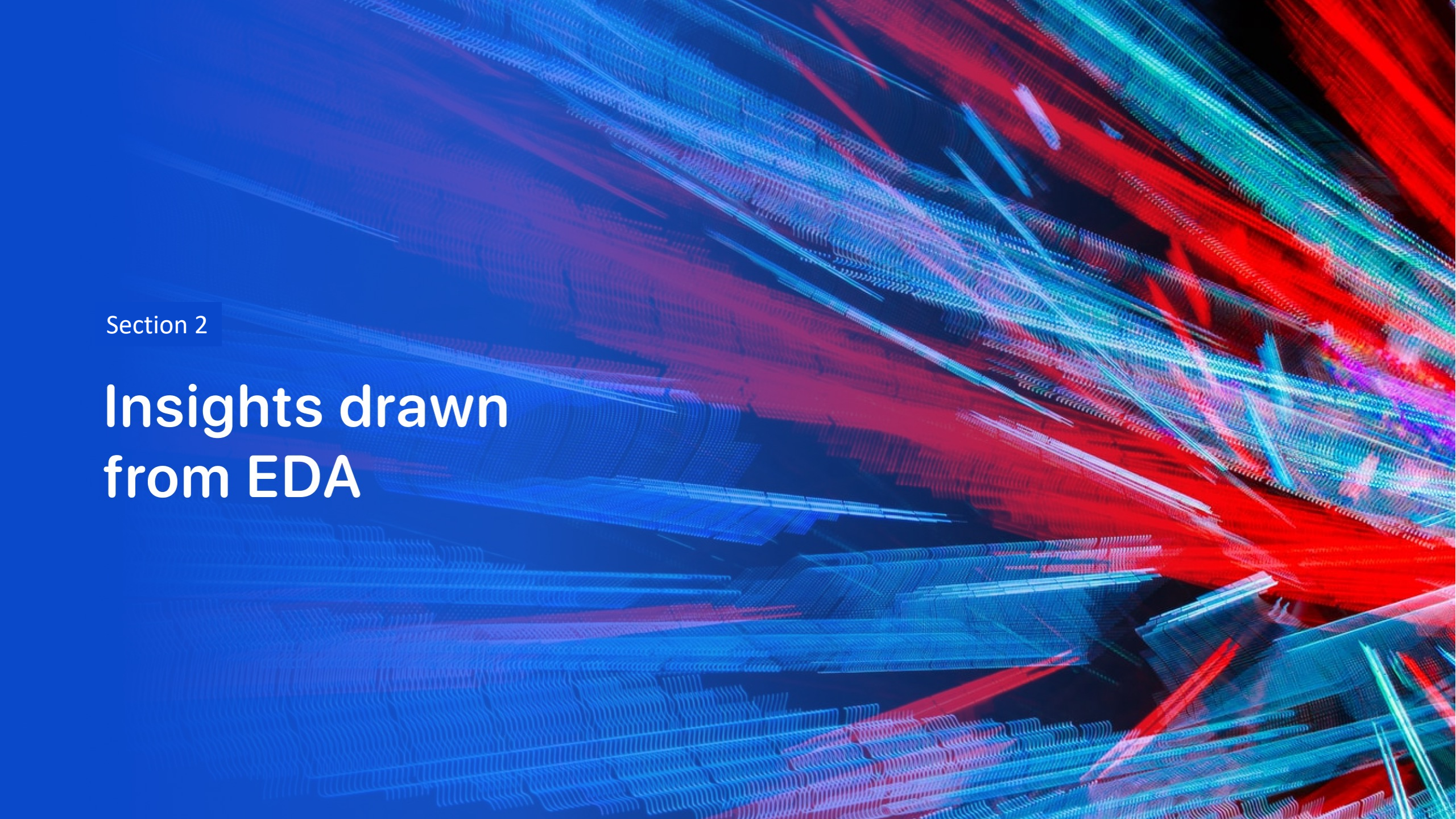
Predictive Analysis (Classification)



https://github.com/mingluzeng/spaceX_assignment/blob/main/SpaceX_Machine Learning Prediction_Part_5.ipynb

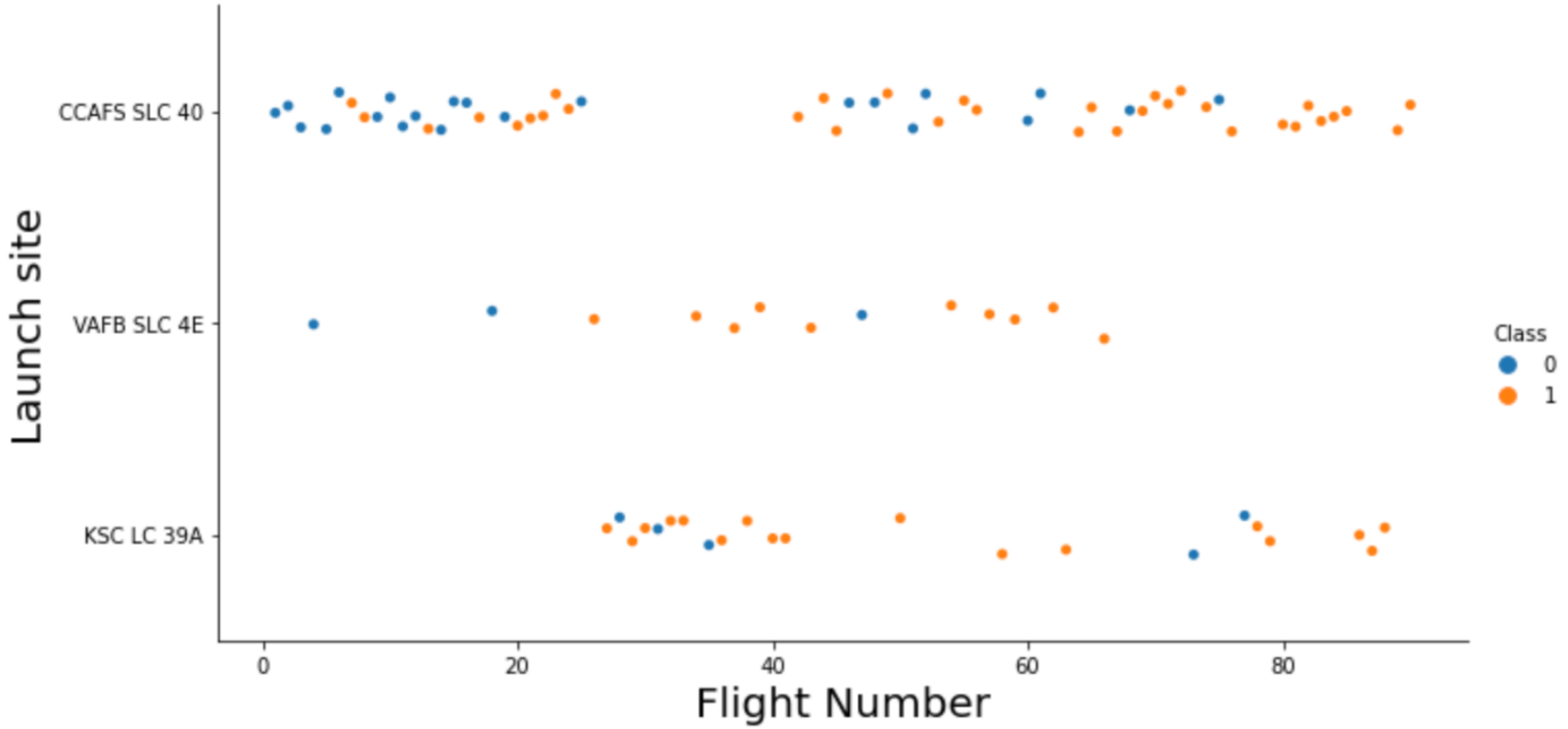
Results

- Each launch site has different successful rate. And KSC LC-39A has the highest successful rate which is 76.9%.
- The successful rate of lighter Payload is higher than heavier Payload
- ES-L1, GEO, HEO, SSO have the highest success rate
- The successful rate increased significantly in 2013
- Decision Tree Classifier has the highest score among four type machine learning


Section 2
Insights drawn from EDA

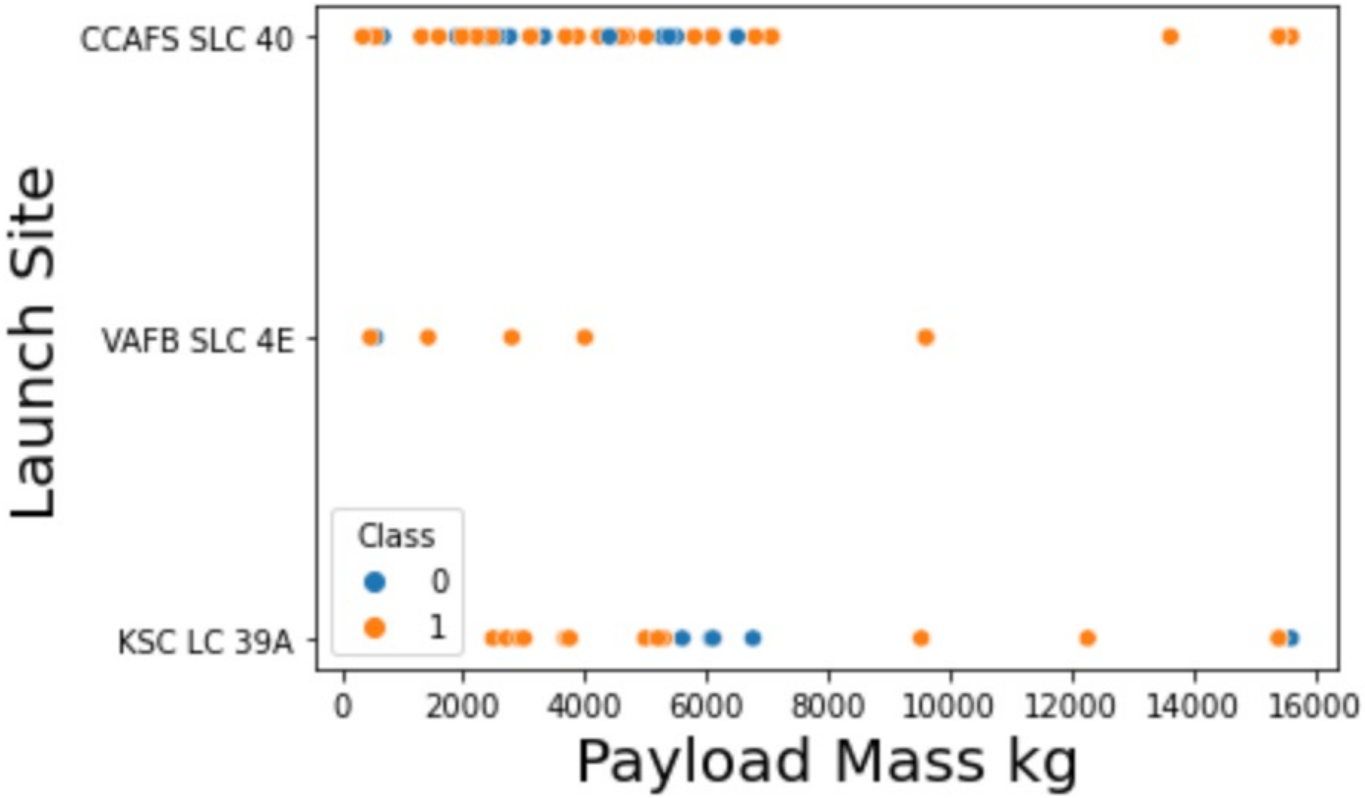
Flight Number vs. Launch Site

- Launch site of CCAFS SLC 40 is significantly higher than others
- The larger the flight amount at a launch site, the greater the success rate at a launch site.



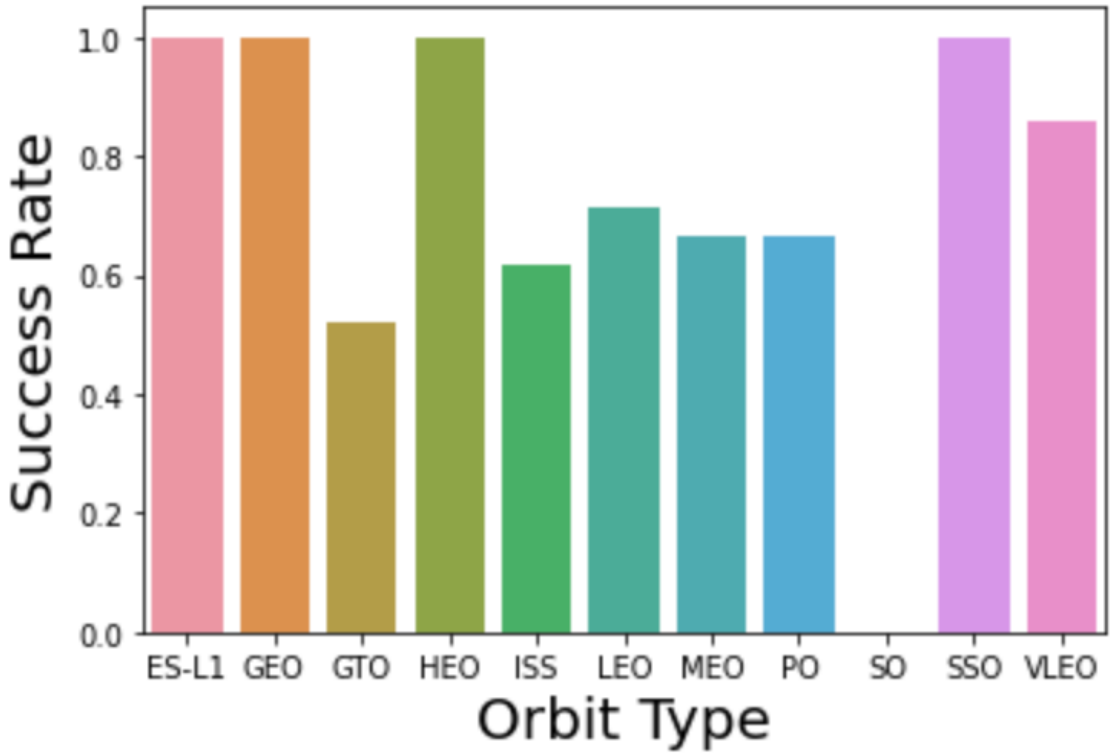
Payload vs. Launch Site

- The Majority of loads with lower mass are launched from CCAFS SLC 40
- There is no heavy payload(more than 1000KG of VAFB SLC 4E)



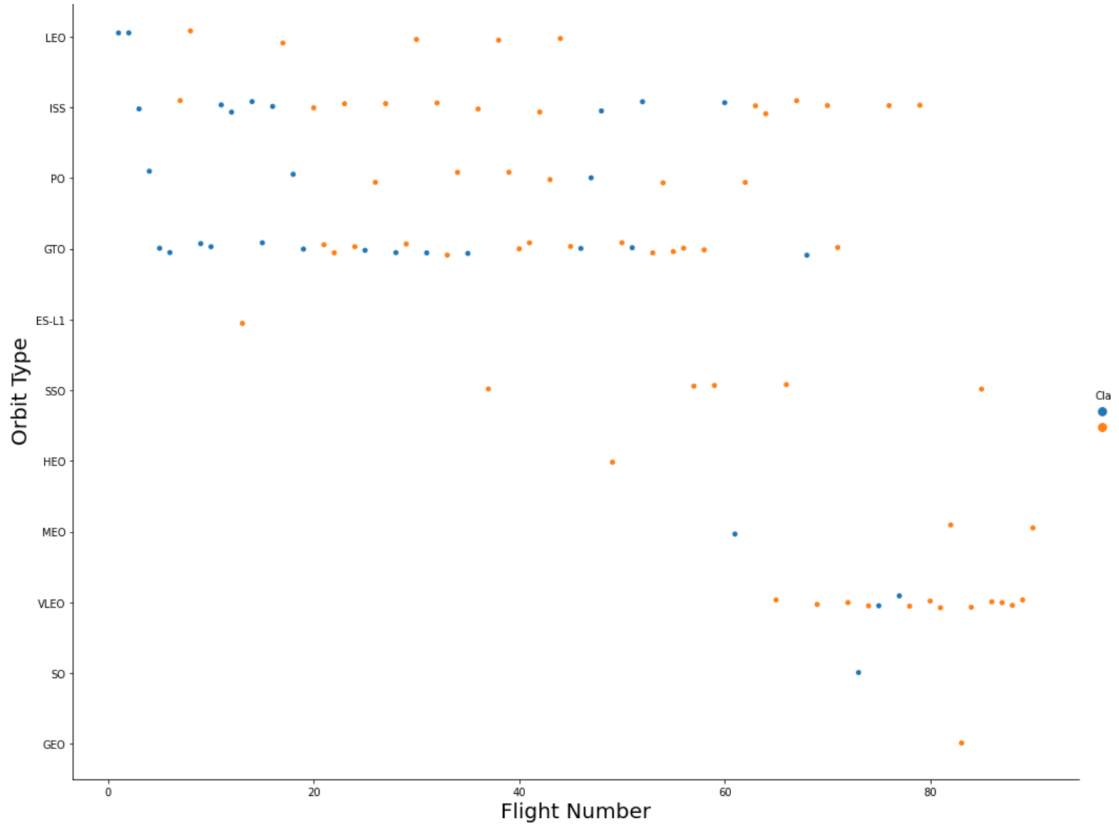
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO have the highest success rate



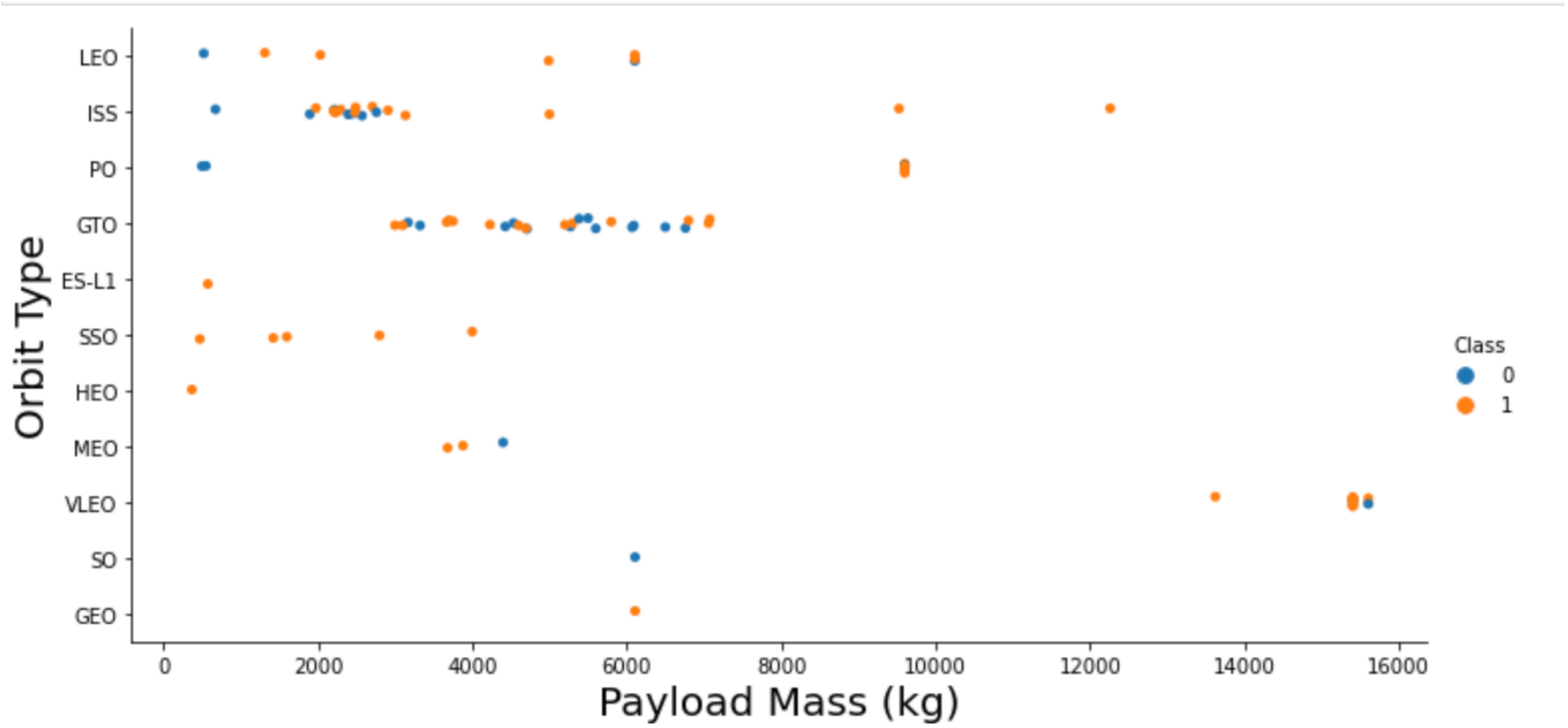
Flight Number vs. Orbit Type

- It displayed successful rate is related to the number of flights in the LEO orbit, and there is no obvious relationship between flight number and the orbit in the GTO orbit



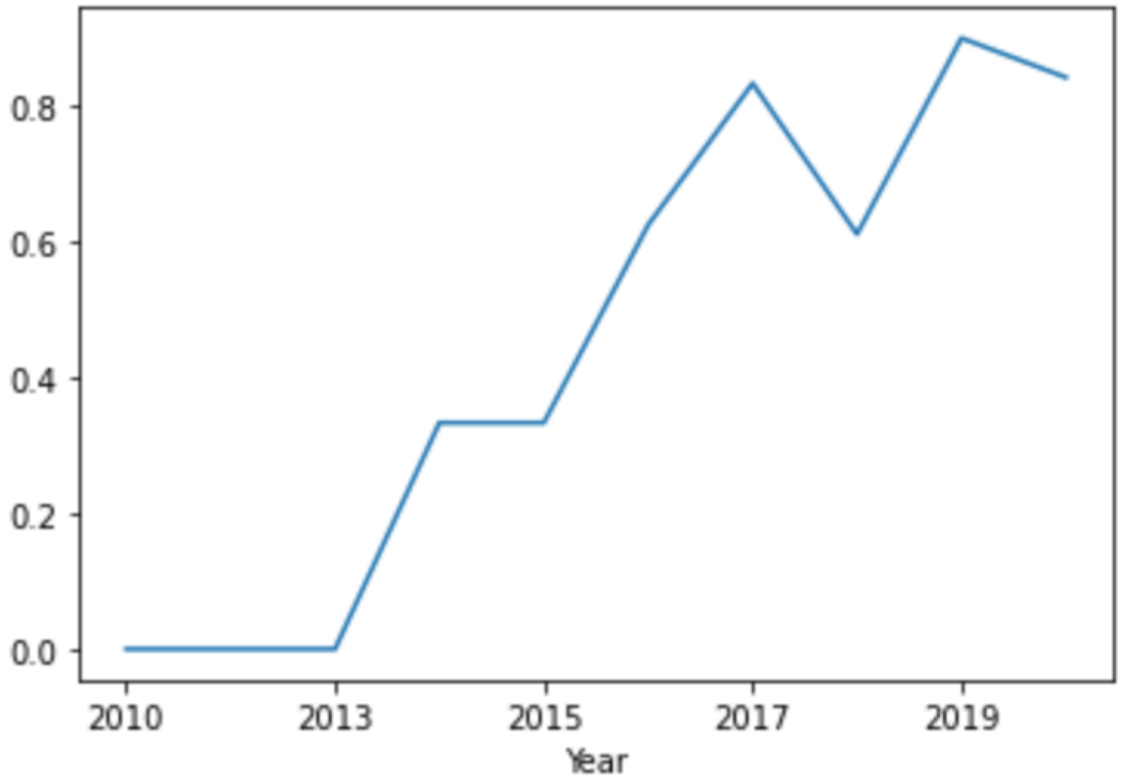
Payload vs. Orbit Type

- We can observe the strong correlation between ISS and Playload from 2000 to 3000, and GTO from 3000 to 8000



Launch Success Yearly Trend

you can observe that the Success rate since 2013 kept increasing till 2020



All Launch Site Names

- There are 4 launch sites

```
[9]: %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

[9]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- **Display 5 records where launch sites begin with the string 'CCA'**

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	

Total Payload Mass

- **Display the total payload mass carried by boosters launched by NASA (CRS)**

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- **Display average payload mass (kg) carried by booster version F9 v1.1**

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- **The date when the first successful landing outcome in ground pad was 2015-12-22.**

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- **List the total number of successful and failure mission outcomes**

```
]: %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * sqlite:///my_data1.db
Done.
```

]:

Mission_Outcome	OUTCOME
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- **The names of the booster_versions which have carried the maximum payload mass.**

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- **The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015**

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

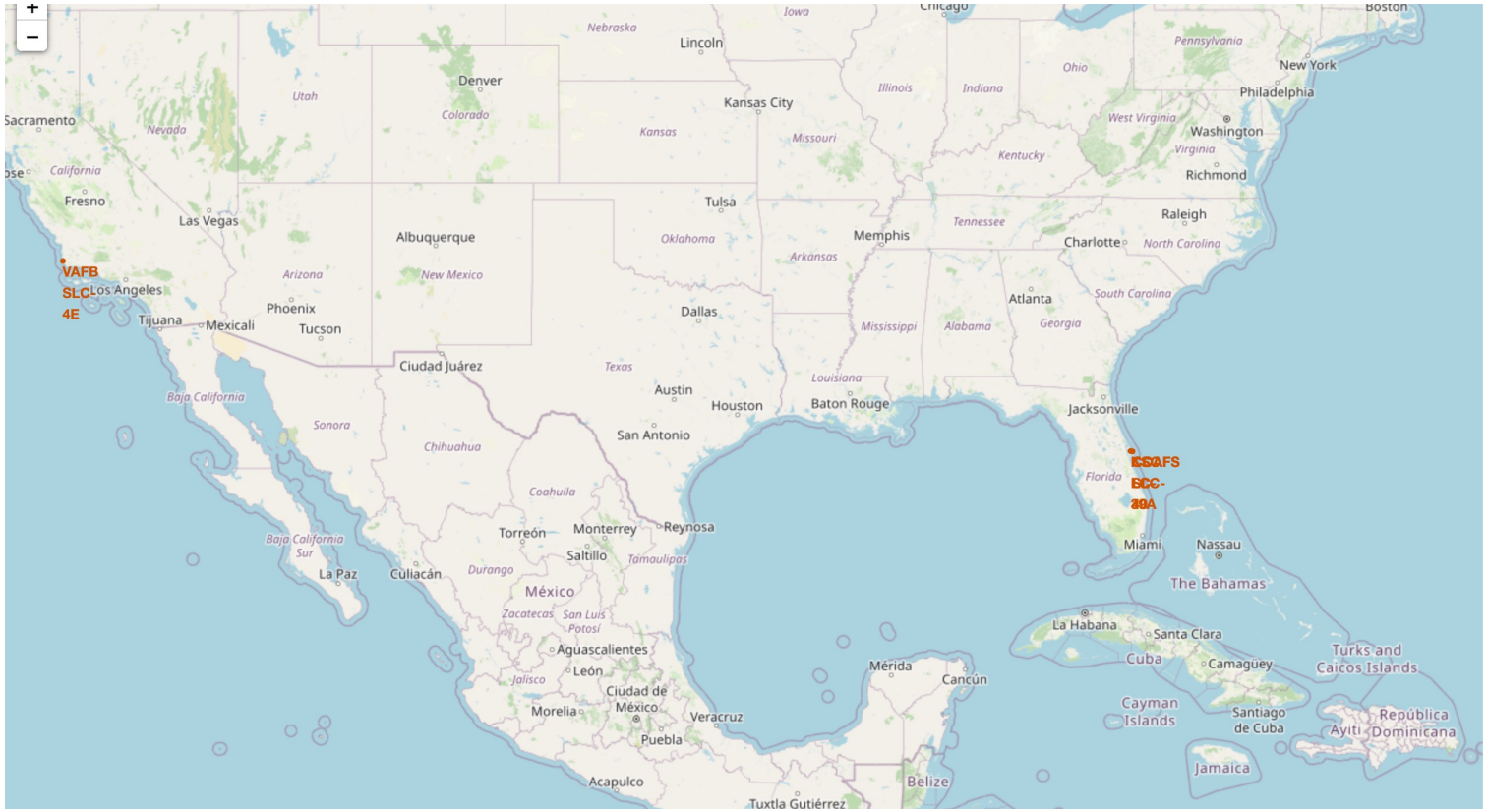
- **Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order**

landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

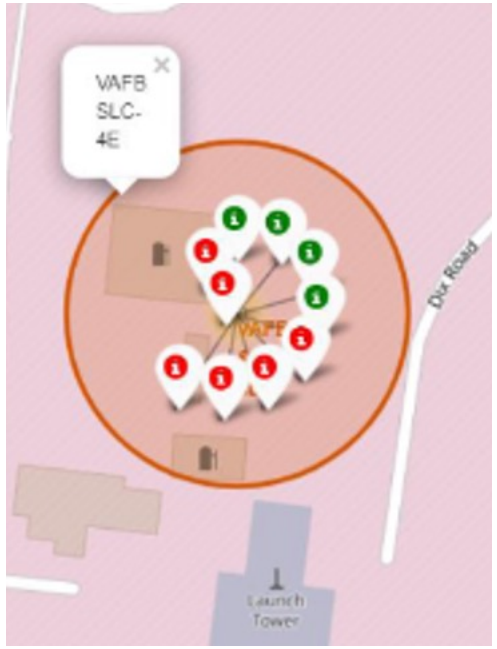
All Launch sites

- launch sites are often located off the coast

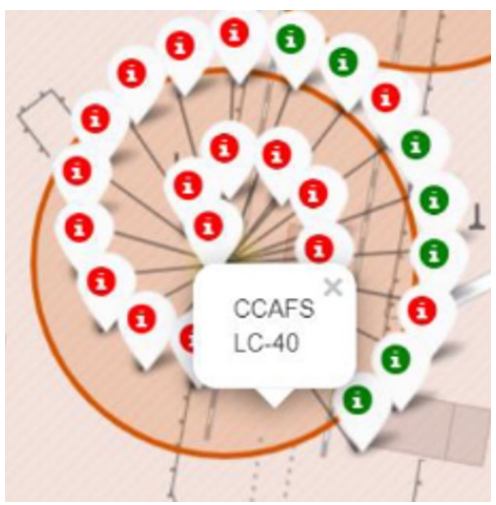


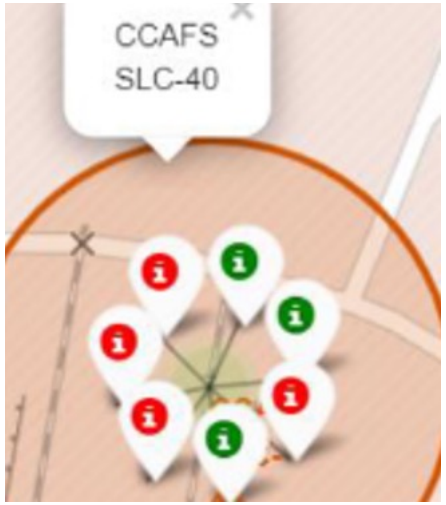
The success/failed launch sites

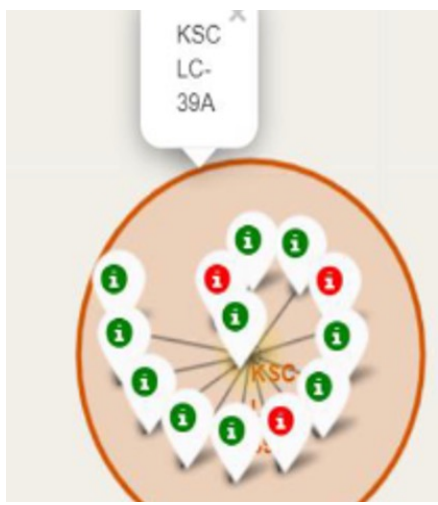
- The green markers means successful, and the red markers means failed



California Launch site



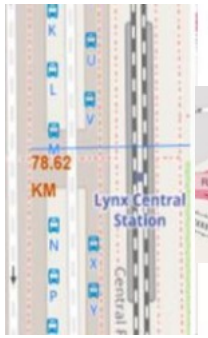




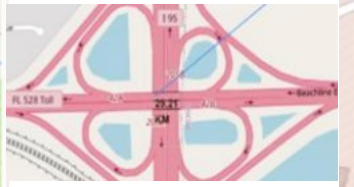
Florida Launch sites

The Distances Between a Launch Site to Its Proximities

- Launch sites are in close to coastline, and keep certain distance away from railways, highway and cities.



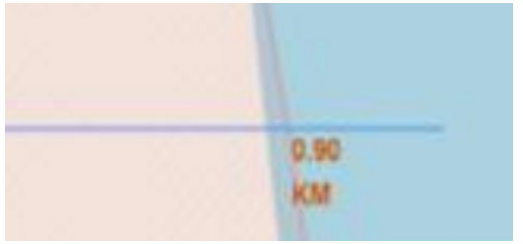
railway



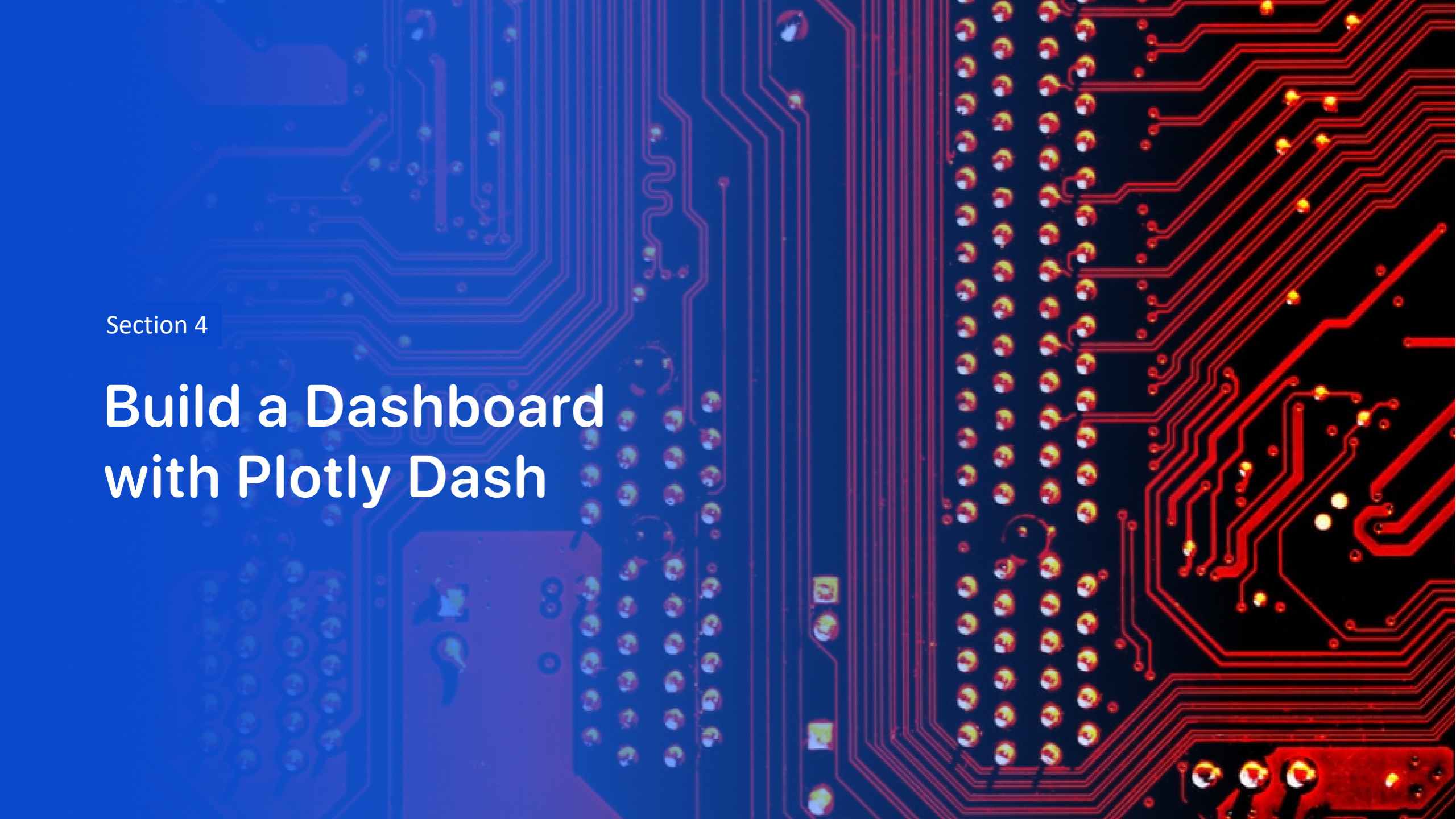
highway

Coast

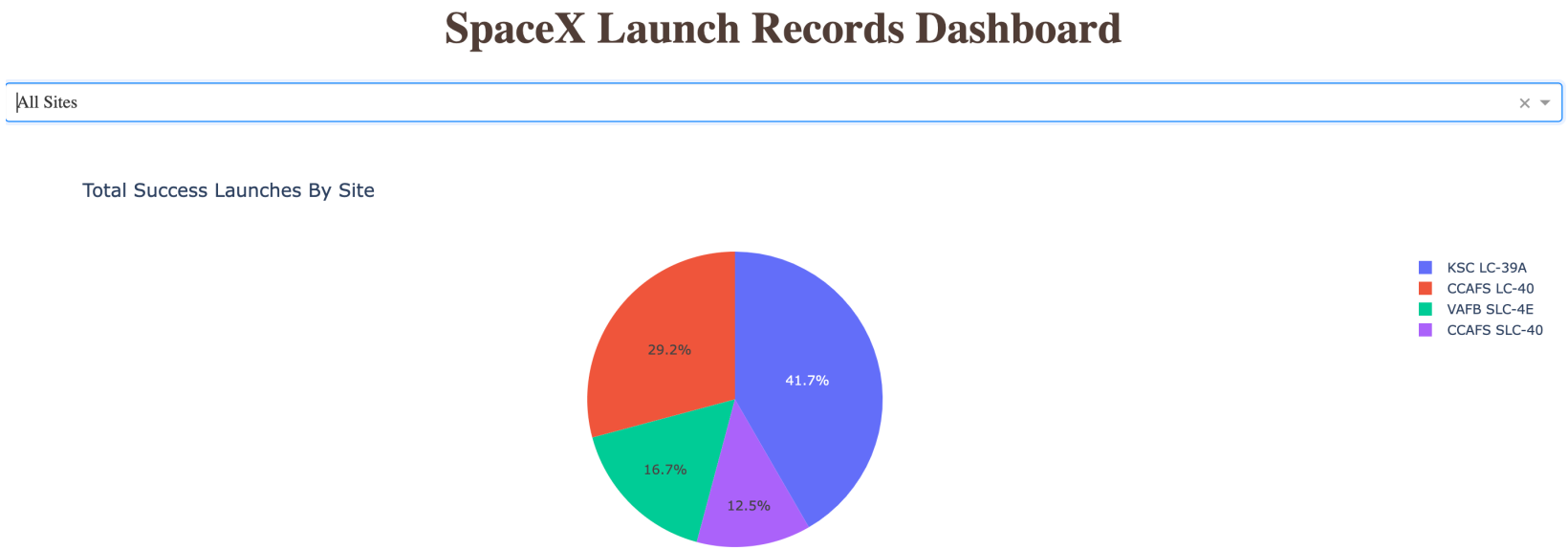
City



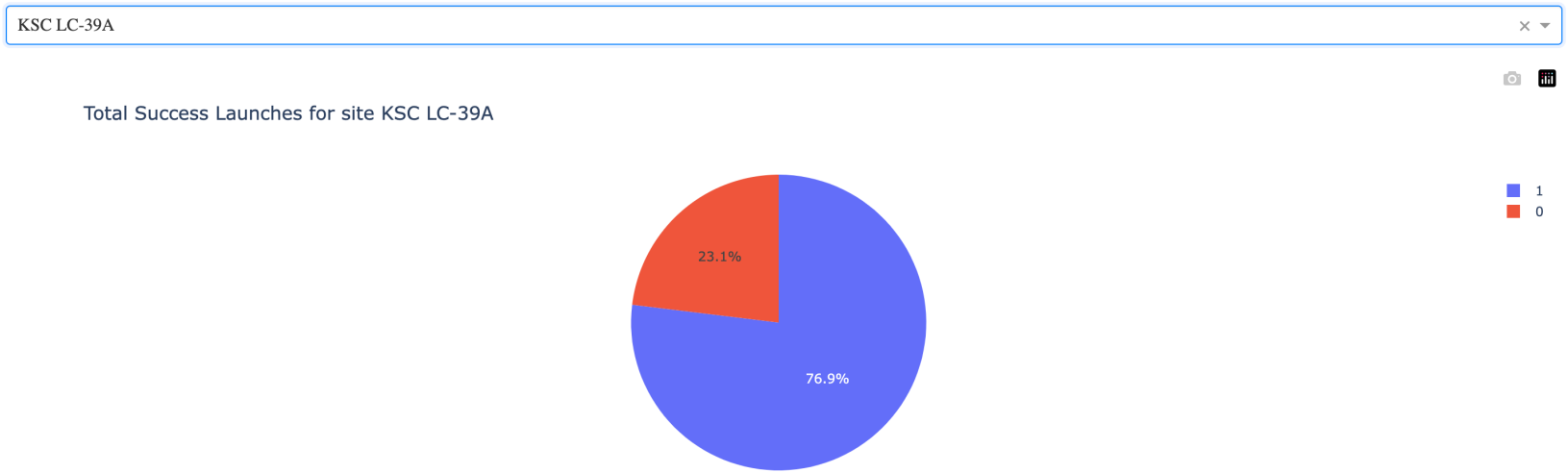
Coastline


Section 4
Build a Dashboard with Plotly Dash

The site with the highest percentage of successful launches is KSC LC-39A(41.7%), and the with the lowest percentage of successful launches is CCAFS SLC-40(12.5%)



KSC LC-39A has the highest successful rate which is 76.9%



Payload vs launch outcome

The successful rate of lighter Payload is higher than heavier Payload

From 4000KG to 10000KG

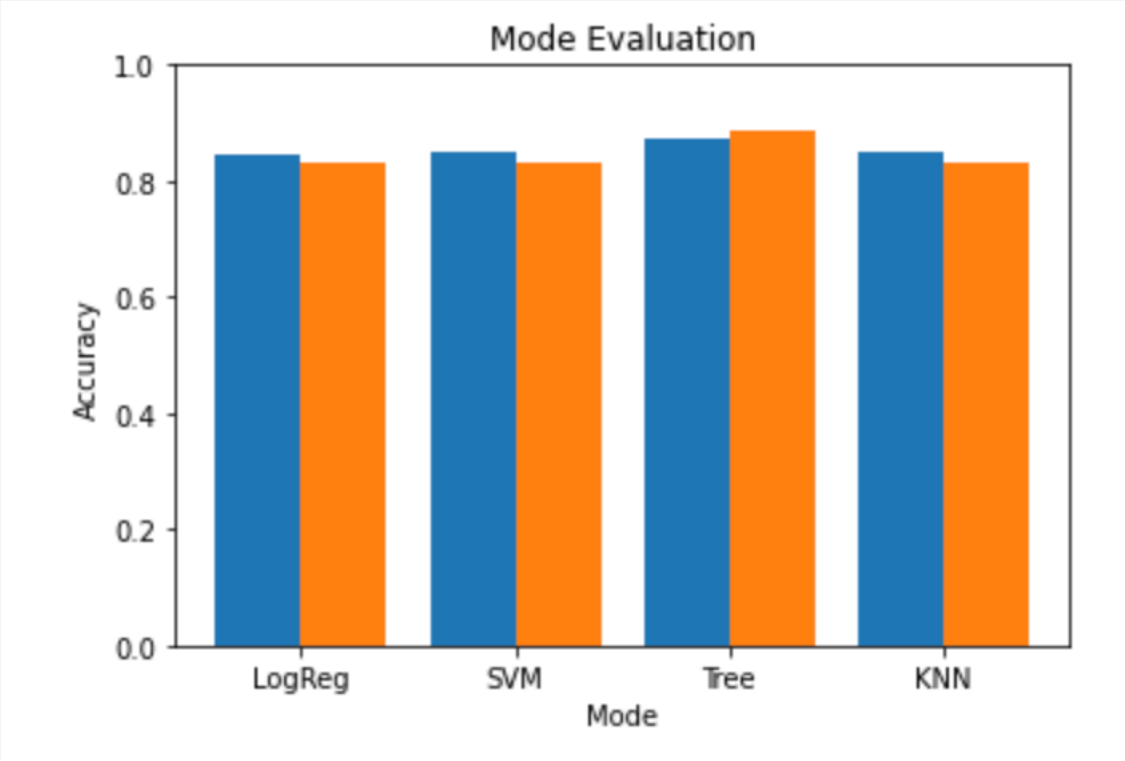
From 0KG to 4000KG

Section 5

Predictive Analysis (Classification)

Classification Accuracy

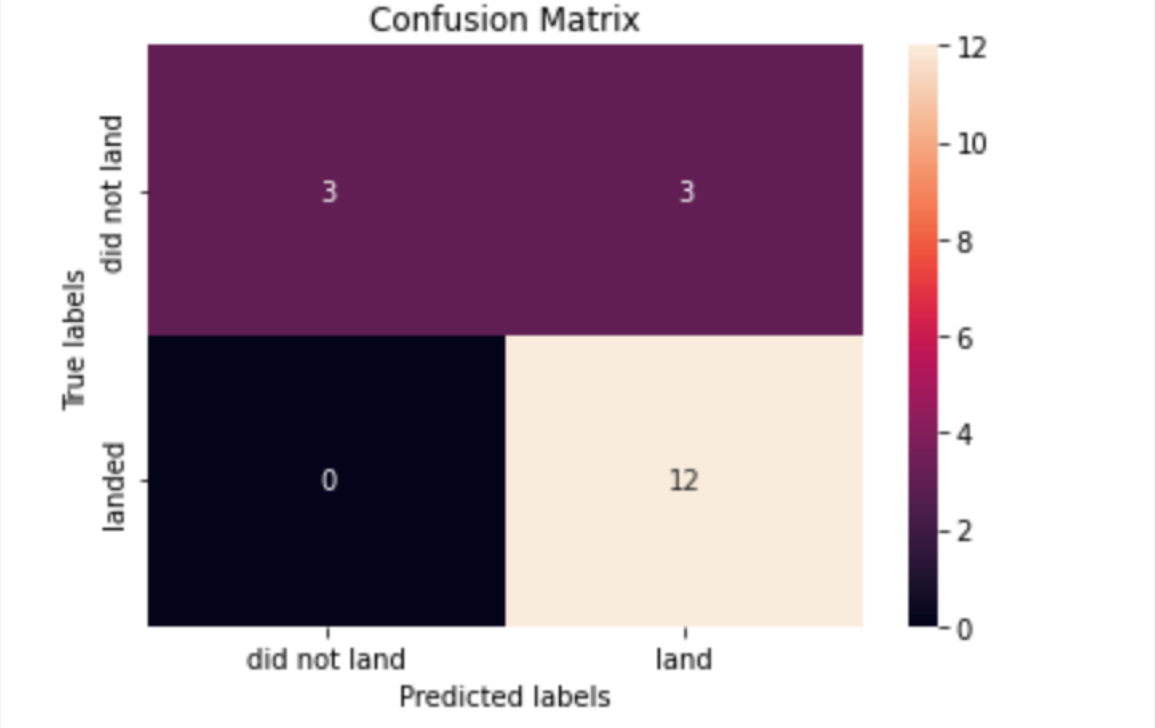
- Decision tree mode has the highest classification accuracy



Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.88889
KNN	0.84821	0.83333

Confusion Matrix

- It is the Confusion matrix of Decision Tree. It proves accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Decision tree mode has the highest accuracy
- The site with the highest percentage of successful launches is KSC LC-39A(41.7%)
- Launch success rate increased significantly in 2013 till 2020
- ES-L1, GEO, HEO, SSO have the highest success rate
- The successful rate of lighter Payload is higher than heavier Payload
- The larger number of flights at a launch site, the great successful rate for the next time of the same launch site

Thank you!